Inventing A Robust Scheme For Passage Lines Histogram Using Indian Languages

N.VISHWANATH
Associate Professor, Dept of ECE
St. Peter's Engineering College
Hyderabad, T.S, India

I.SUMATI
M.Tech Student, Dept of ECE
St. Peter's Engineering College
Hyderabad, T.S, India

Abstract: **Most of the Indian scripts are originated from Brahmi script through various transformations. Writing style of the Indian scripts considered in this paper is from left to right, and concept of upper/lower case is absent in these scripts A complete line and word segmentation system for some popular Indian printed languages is presented here. In the present technique subsequent to getting separating lines, it should be checked whether each separating line passes through the white gap between two consecutive lines or it crosses some components of text lines. Document image binarization is an useful method to convert a gray image into two tone. Both foreground and background information are used here for accurate line segmentation. We have used histogram based properties to binarize the documents taken as a data set. The digitized text images are first converted into two-tone images using a histogram based thresholding approach. Our method can take care of this situation accurately. We have tested our method on documents of Bangla, Devnagari, Kannada, Telugu scripts as well as some multi-script documents and we have obtained encouraging results from our proposed technique. There may be some touching or overlapping characters between two consecutive text lines and most of the line segmentation errors are generated due to touching and overlapping character occurrences. Sometimes, interline space and noises make line segmentation a difficult task.**

Keywords: **Document Image Binarization; Optical Character Recognition; Document Analysis; Hough Transform;**

I. INTRODUCTION

Research in OCR is popular for its various application potentials in banks, library automation post-offices and defense organizations. Character recognition as an aid to the visually handicapped was at first attempted by the Russian scientist Tyurin in 1900. The OCR technology took a major turn in the middle of 1950s with the development of digital computer and improved scanning devices [1]. As a conventional technique for text line segmentation, global horizontal projection analysis of black pixels has been utilized. Currently, PC-based systems are commercially available to read printed documents of single font with very high accuracy and documents of multiple fonts with reasonable accuracy. Line and word segmentation is one of the important step of OCR systems. In this paper we have proposed a robust method for segmentation of individual text lines based on the modified histogram obtained from run length based smearing. These techniques may be categorized into three groups as follows: (i) Projection profile based techniques, (ii) Hough transform based techniques, (iii) Thinning based approach. The positions of potential piece-wise separating lines are obtained for each stripe using partial horizontal projection on each stripe. The potential separating lines are then connected to achieve complete separating lines for all respective text lines located in the textpage image [2]. In this paper we have proposed a robust method for segmentation of documents info lines and words and the method is based on the modified histogram obtained from run length based smearing. In word segmentation method, a text line has taken as an input. After a text line is segmented, it is scanned vertically. If in one vertical scan two or less black pixels are encountered then the scan is denoted by 0, else the scan is denoted by the number of black pixels. Foreground and background information is also used for accurate line segmentation.

Fig.1.Illustration of overlapping

II. METHODOLOGY

Most of the Indian scripts are originated from Brahmi script through various transformations. Writing style of the Indian scripts considered in this paper is from left to right, and concept of upper/lower case is absent in these scripts. Devnagari is the most popular script in India and the most popular Indian language Hindi is written in Devnagari script. Bangla, the second most popular language in India and the fifth most popular language in the world, is an ancient Indo-Aryans language. Telugu is the 3rd most popular scripts in India. It is the official language of the southern Indian state, Andhra Pradesh [3]. Telugu

is also spoken in Bahrain, Fiji, Malaysia, Mauritius, Singapore and the UAE. Document image binarization is an useful method to convert a gray image into two tone. Global binarization and locally adaptive binarization are two popular types of binarization methods. There are few categories of binarization methods, such as Histogram-based, Clustering-based, Entropy-based, Object attribute-based, Spatial binarization and Locally adaptive etc. We have used histogram based properties to binarize the documents taken as a data set. The digitized text images are first converted into two-tone images using a histogram based thresholding approach. Skew estimation and correction are important preprocessing steps of line and word segmentation approaches [4]. It deals with skew estimation of a class of scripts that includes some major Indian Languages like Bangla, Devnagari, Kannada, and Telugu. In this work, we use a Hough transform based technique for skew angle estimation. To reduce the amount of data to be processed by the Hough transform, we compute some candidate points considering some selected components from the image. There are several steps in the line segmentation method: i) In this step we use run length smearing technique to increase the strength of the histogram.ii) Getting the histogram of every line from the smoothed document page, we consider the highest peak among all the peaks of the horizontal projection profile. We draw the horizontal lines based on this middle point of the width of histogram. In some cases all peak of histograms do not cross this vertical line [5]. For these cases we find distances between middle lines and find the average value of these distances.iii) from the starting point of first histogram we vertically scan the region in between the first middle and second middle line of histogram until we get first two white pixels. To remove errors we use the above technique for candidate line separator and this technique is useful for all scripts like Bangla, Devnagari, Telugu and Kannada scripts considered here. iv) In a text-page, either overlapping or touching or both the problems of overlapping and touching may occur in many positions of two consecutive text lines in the text-page. In the present technique subsequent to getting separating lines, it should be checked whether each separating line passes through the white gap between two consecutive lines or it crosses some components of text lines. Based on a statistical study, it was explored that when two components touched, in most of the cases the positions with minimum stroke width within this zone happened to be the position of touching of two components [6]. In word segmentation method, a text line has taken as an input. After a text line is segmented, it is scanned vertically. If in one vertical scan two or less black pixels are encountered then the scan is denoted by 0, else the scan is denoted by the number of black pixels. We have tested the algorithm on single script and multi-script documents.

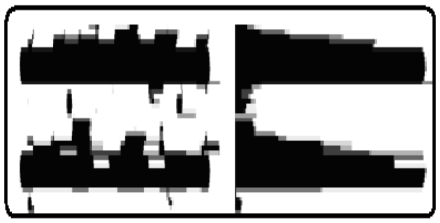

Fig.2.Histogram model

III. CONCLUSION

To take care of the problem of overlapping, the contour points of the component are traced. The intersection point of the separating line and the component is considered as starting point for contour traversal. Touching generally occurs in this zone. Line and word segmentation is one of the important steps of OCR systems. Document image binarization is an useful method to convert a gray image into two tone. Global binarization and locally adaptive binarization are two popular types of binarization methods. In Optical Character Recognition (OCR), the text lines in a document must be segmented properly before recognition. In this paper we have proposed a robust method for segmentation of individual text lines based on the modified histogram obtained from run length based smearing. To remove errors we use the above technique for candidate line separator and this technique is useful for all scripts like Bangla, Devnagari, Telugu and Kannada scripts considered here. Both foreground and background information are used here for taking care of touching characters for accurate segmentation. In Optical Character Recognition, the text lines in a document must be segmented properly before recognition. Correctness/Incorrect-ness of text line segmentation directly affects accuracies of word/character segmentation and consequently changes the accuracies of word/character recognitions.

IV. REFERENCES

[1] U. Pal and Sagarika Datta, "Segmentation of Bangla Unconstrained Handwritten Text", Proc. 7th Int. Conf. on Document Analysis and Recognition, pp.1128-1132, 2003.

[2] H. Yan, “Skew correction of document images using interline cross-correlation”, CVGIP: Graph. Models Image Process, vol. 55, pp. 538–543, 1993.

[3] L. O’Gorman, “The document spectrum for page layout analysis”, IEEE Trans. Pattern Anal. Mach. Intell., vol.15, pp. 1162–1173, 1993.

[4] D. S. Le, G. R. Thoma, and H. Wechsler, "Automatic page orientation and skew angle detection for binary document images", Pattern Recognition, vol. 27, pp.1325-1344, 1994.

[5] Vijay Kumar, Pankaj K.Senegar, "Segmentation of Printed Text in Devnagari Script and Gurmukhi Script ", IJCA: International Journal of Computer Applications, Vol.3,pp. 24-29, 2010.

[6] U. Pal and B.B. Chaudhuri, "Indian script character recognition: A Survey", Pattern Recognition, vol. 37, pp. 1887-1899, 2004.

